TRUST AND ECOLOGICAL RATIONALITY IN A COMPUTING CONTEXT[9]

Jeff Buechner
Department of Philosophy, Rutgers University-Newark
buechner@rci.rutgers.edu
The Saul Kripke Center, CUNY, The Graduate Center
JBuechner@gc.cuny.edu

Abstract
In this paper, I examine a key issue affecting trust in the context of a computing environment, as it affects human agents and artificial agents. Specifically, the paper focuses on the role that "resource conservation" plays in an analysis of moral trust and epistemic trust involving agents. I will argue that resource conservation is a necessary condition in the definition of a moral trust relation, that there is a conceptual relationship between a moral trust relation and epistemic trust —that epistemic trust is a necessary condition for moral trust—, and that a moral trust relation might provide a basic mechanism for the realization of ecological rationality in human agents.

Keywords
Frame problem, ecological rationality, epistemic trust, moral trust, resource conservation

Introduction

My project is to examine a surprising aspect of the trust relation that has been neglected in the literature—the ways in which a trust relation can conserve valuable computational resources. I will argue that this aspect of the trust relation is a necessary aspect; thus, relations in which it is missing are not "authentic trust relations." Moreover, the nature of the resource properties conserved varies in different kinds of trust relation. In particular, in e-networks, two different kinds of trust relation are at work—a moral and an epistemic trust relation. They are often in considerable tension, especially with respect to the conservation role of trust relations. However, establishing knowledge claims also requires trust among agents (on a social conception of knowledge). In some cases, epistemically trusting other agents may preclude morally trusting those same agents, and conversely. The interactions between the two trust relations can be explained in terms of conserving their resource properties.

 This is a revised version of a paper presented at the Ninth International Conference on Computer Ethics: Philosophical Enquiry and included in the conference proceedings.

In this paper I will articulate several resource-conserving features of trust relations in general, as well as elaborate on the connection between ecological rationality and trust relations. In particular, I will argue that in the concept of a moral trust relation there is a necessary component in the form of an epistemic trust relation. I will describe the formal, iterative structure of agent-agent trust interactions that serves to non-circularly establish the trust relation, to reinforce it, and to bootstrap its strength. The major focus of this paper, though, is to defend the claim that an essential feature of the trust relation is resource conservation.

Frame problems, resource conservation, and successful navigation of the world

Human agents are able to perform reliably well in executing fairly mundane plans and actions in the everyday world—crossing the street without being hit by a car, buying groceries, making sandwiches, and so on. However, computational descriptions of these activities are bewilderingly complex, and algorithms for executing them are computationally intractable. Many of these plans and actions fall under the rubric of 'frame problems.'[1] The general form of a frame problem is that to achieve reliability in reaching some goal, an agent must do α, but to do α requires an infeasible amount of valuable computational resources—β—which the agent does not have. The problem is how the agent can achieve reliability in reaching their goal without using β resources (and so, without doing α). There is an analogy with problems in computational complexity: how can problem X be solved without using β resources, where solving X requires —because it is the complexity profile of the problem—β resources? In cases where a computational system does not have the resources to solve a problem requiring those resources, what sorts of solutions are available to the agent, and how reliable are they? Some solutions allow a computational system to do all of the steps necessary for solving X without expending the resources (such as DNA computers), while other solutions relax somewhat the requirements on reliability. Examples of these include probabilistic solutions and approximation solutions to X.

Although there are several different kinds of frame problems, all of them have the general characteristic sketched above: achieving reliability in reaching one's goals (of such-and-such a kind) appears to be computationally intractable. However, humans are able to achieve reliability and achieve their goals (of such-and-such a kind) in many different kinds of situation (though certainly not all) without doing what is computationally intractable. How is it done?

Trust and resource conservation

There is a good deal in the literature on the concept of moral trust that suggests that one function of the trust relation is to provide a way of sidestepping a computationally intractable problem, though no one has explicitly claimed that such a feature is an essential feature of the trust relation, let alone explicitly recognizing it. Annette Baier[2] has made the claim that in a trust relation we entrust something—a thing, a task, a psychological feature—to the care of another. Where A cares about something, and A trusts B, A trusts that B will care about it as well. However, in order for B to care about something that A cares about, both A and B must exhibit a certain form of discretion. That discretion is necessary for proper caring. Indeed, Baier claims, especially in her essay "Trust and Anti-trust," that it is not feasible to explicitly articulate, in advance, just exactly how the caring—on both sides of the trust relation—is to be administered. If both agents had to know the specific details of how caring proceeds in all possible circumstances before they could engage in a trust relation, the trust relation could not take place, since it would be a computationally infeasible task to acquire that knowledge. Notice that it is part of the trust relation that explicit listing of the conditions for caring in all possible circumstances is not needed.

Baier's account of the trust relation in terms of caring has been criticized by Margaret Urban Walker as not saying enough about those aspects of caring that cannot be explicitly articulated in advance.[3] Walker thinks that it is also not possible to articulate in advance what one is entrusted to care about. Whether Walker is right about this, though, is not essential to my point, since even if it is only providing for the caring relation that cannot be explicitly articulated in advance, it is still the case that the trust relation allows one to have a relation where a computationally infeasible task, that might be thought to be necessary for the trust relation, cannot be performed. What has not been noticed in the literature on trust is that this very feature —the trust relation renders superfluous a feature that might be thought necessary to establish a trust relation—should be taken to be an essential feature of a trust relation.

Walker claims that "… what can we say we are entrusting to all others when any of us walks down the street without concern? Here there is not one norm presupposed but indefinitely many, any of which might be identified only in the breach or under threat, and each of these norms requires many things we cannot enumerate in advance."[4] If we had to enumerate in advance all of these features in order for the trust relation to be satisfied, we would not be able to do so either optimally or even suboptimally, since to do so would require more resources than we

have. However, if it is a feature of trusting the other that we trust them—and they trust us—to do the right things when the circumstances demand that, then it is certainly an essential property of trusting that we are able to avoid having to perform tasks that require too many resources to be satisfactorily performed.

A notion of trust put forth by Trudy Govier has several features, one of which is that an agent who trusts another agent has a disposition to provide a positive interpretation of the actions of those whom that agent entrusts.[5] This is an important way of saving valuable resources, since it would be difficult for any given agent to continually monitor the actions of another agent in the service of evaluating them with respect to several different dimensions of evaluation. Govier also thinks that entrusting agents in a trust relation attribute a sense of general integrity to entrusted agents. This, too, is a means of saving valuable resources, since it would be difficult, if not impossible, for an agent to continually re-evaluate another agent as a whole with respect to integrity. The difficulty is that integrity is a capacity of an agent—in the sense of being a virtue —as much as it is a disposition to behave in a certain, morally evaluable way. Consequently, monitoring it is difficult, since it will, for example, fail in some contexts or be difficult to recognize in other contexts.

Govier also brings the notion of expectation into the picture of trust, but in an interesting way—she takes it to be a feature of trust that trusting agents have expectations that entrusted agents will exhibit good behavior, where the trusting agent's expectations are based solely on beliefs about the entrusted agent's motivation and competence. Here, too, resources are conserved, since having an expectation that an agent will behave in a certain way sidesteps the need to actually observe the agent's behavior from time to time (or at certain times reflecting a good statistical sample). It also sidesteps having to explicitly enumerate and watch for multiple specific kinds of behaviors, especially aberrant ones. The idea that an expectation plays such a role in the trust relation has never been explicitly developed, but it is clear that it is central to the notion of trust. Moreover, although positive expectations of behavior on the part of the entrusted do not singly constitute the trust relation, it is important to see that those expectations—and the role they play in conserving resources an agent possesses—are justified when the normative notions that flesh out the notion of trust are brought into play.

Govier also sees a need to bring into the trust relation the idea that a trusting agent accepts that there is risk in trusting and that this risk leaves the agent vulnerable in certain ways.

However, it is not the case that the trusting agent actually computes a risk profile, or provides an exhaustive list of the ways in which she is vulnerable. Rather, the trusting agent merely acknowledges that she is at risk or is vulnerable, without developing additional justifications for this contention by listing the ways in which that can occur. Indeed, it would be inimical to the trust relation if the trusting agent *did* provide risk profiles and exhaustive enumerations of vulnerabilities.

The idea that in the trust relation there are constraints on interpreting the actions of entrusted agents has been considerably developed by Karen Jones, who focuses on the ways in which unfavorable interpretations of another's actions are jettisoned and, in their place, favorable interpretations offered.[6] Moreover, the range of favorable interpretations is kept quite small, so that the entrusting agent will not have to survey different kinds of favorable interpretations—and, presumably, ascribe to the entrusted agent various properties based on that interpretation. In this way, an agent saves valuable resources. Obviously, there are circumstances in which an entrusting agent must give up a favorable interpretation. However, these circumstances are those where the behavior of the entrusted agent is out of the ordinary, or abnormal. Just as in frame problems in artificial intelligence, there are circumstances where the typical behavior of an agent (or a process, etc.) fails to occur, so, too, there are circumstances in trust relations where an entrusted agent acts abnormally (with respect to the normative expectations of the entrusting agent). It is a hard question in non-monotonic reasoning (which provides a formal class of solutions to frame problems) as to how an agent is able to detect such abnormalities without having to explicitly list all normal and abnormal conditions and then monitor each situation with respect to this list for evidence of an abnormality. The correlative question for a trust relation is how agents engaged in such relations are able to determine when the relation should be broken—that is, are able to determine when conditions are out of the ordinary. If agents cannot tell when conditions are out of the ordinary, it is not clear that they fully understand what it is to commit to a trust relation.

Judith Baker has argued that entrusting agents may believe that an entrusted agent is worthy of trust, even where there is evidence that this is not so.[7] This is a difficult topic that is important—since there is a relation between epistemic trust and moral trust—and this point has not received the attention it deserves. Where agents do not rely upon the best evidence in making decisions, they can be accused of acting irrationally in an epistemic sense. But can they

be accused of acting irrationally in a moral sense? This is not a case of epistemic akrasia (where an agent acts irrationally in the view of an epistemic norm, even though, in the view of another, less binding epistemic norm, the agent acts rationally); rather, it is an instance of "hybrid akrasia," where the agent acts irrationally in the eyes of an epistemic norm, but rationally in the eyes of a moral norm. The point, though, is that even where there is evidence against trusting another agent, one might trust that agent nonetheless. This, in turn, shows that there are regions in the trust relation where it is not clear whether an agent should trust or not trust another agent, but that the agent should, unless there are clear indications to act otherwise, trust the other agent. Thus, such aspects of the trust relation are second-order aspects—that is, the agent knows that she is in a region where trust might be given up since there is epistemic evidence that the entrusted cannot be trusted. But, reflecting on the trust relation, she affirms her trust in the entrusted agent on the grounds that it is part of the trust relation to trust the other agent even where there are certain grounds for giving it up (though the grounds cannot be conclusive, whatever that might mean in this context).

Frame problems arising in moral theories

There is little reason to think that the kinds of moral interactions between human beings are, though normatively constituted, immune to the complex contingencies of the everyday empirical world. It is the complexity of these contingencies that creates frame problems, and so one would think that some sort of frame problem arises for moral interactions, since these interactions occur in the real world. Certainly, consequentialist moral theories suffer frame-like problems, in the following way. When an action is performed, it has many consequences in the real world, depending upon the temporal and spatial width of the intervals one considers. But natural physical laws apply to those spatial intervals only if conditions are ideal. A ball will not continue to roll in a certain direction if someone picks it up, if there is something which blocks its path, if the ball implodes, and so on. This is called the qualification problem.[8] The smaller the temporal and spatial intervals, the fewer consequences there are to consider (though the complexity of any consequence can be great even for minimal temporal and spatial intervals). Thus, the smaller the temporal and spatial intervals, the more reliable a prediction will be, but at the cost of efficiency, since one must make many more predictions (the shorter the temporal and spatial intervals). This is called the extended prediction problem.[9]

One question is how a consequentialist moral agent delimits temporal and spatial intervals so that (i) the question is not begged as to whether the intervals arrived at accord with consequentialism (in particular, that one does not delimit temporal and spatial intervals in order to justify a particular moral claim) and (ii) the intervals arrived at do not require for their construction computational resources the agent does not possess. For instance, it might happen that determining a consequence in real time and with real spatial parameters requires making calculations that an agent does not have the time to perform, or the instruments with which to make the calculations. If an agent cannot perform all of the calculations demanded by the consequentialist construal of the situation, owing to resource limitations, and on the basis of the calculations which she can perform, decides to perform some action, what shall we say? That is, if the action is one which is not the optimal consequentialist action, is it the morally correct action? If not, what shall we say about resource limitations? If the morally correct action requires one to do something which one cannot do, then can it be the morally correct action? If it is not, then either resource limitations play a role in determining normatively correct moral actions or they are considered to be features of the environment which prevent one from engaging in normatively correct moral actions.

However, the role of resource limitations and the status of the frame problem in evaluating consequentialist moral theories expose a *problem* in understanding how to use a consequentialist theory, and this problem is one about how norms are applied in a contingent world. It is the presence of this problem which haunts many of the applications of any moral theory to actions which occur in the real world. To generalize, it is a problem which all normative theories, whether moral or not, must address. To comply with norms, one must be able to successfully navigate contingencies in the world, including restrictions on one's resources, especially computational resources. This is not news to anyone, however. But that a normative notion might be a means to *solving* such problems is not known in the literature. This is the surprising connection of trust and the conservation of resources (most notably, computational resources). We will argue that one necessary function of (the concept of) a trust relation is to conserve resources that agents—the trustor and the trustee—would otherwise have to expend in order to accomplish their goals. This important feature of a trust relation has escaped notice in the literature, perhaps because of a myopic fixation on what the trustor and trustee *should* do in a trust relation, and not enough (if any) attention on how this is to happen.

The most obvious point is that in trusting an agent to do some action, you save resources by not having to do that action yourself. Of course, in trusting an agent to do that action, you expend resources. One question is whether the amount of resources saved in trusting an agent to do that action is greater than the resources needed to maintain a trust relation—that the agent does that action. But this focus is incomplete, since there will be world contingencies which that agent will have to address in order to perform the action. Moreover, there are world contingencies that the trustor will have to address in order that her view of the entrusted agent as reliable has some justification. Finally, there are other agents necessary for the means by which the entrusted agent performs the action. These other agents, though not directly part of the trust relation between trustor and the entrusted, must still be trusted in order that the entrusted agent successfully performs the action.

Notice that it would be (i) infeasible for an entrusting agent to make computations of whether the resources saved by having the entrusted agent perform an action outweighs the resources expended by the entrusting agent in doing that action, (ii) infeasible to compute the world contingencies that might happen that need to be addressed by the entrusted agent and the world contingencies that might happen that need to be addressed by the entrusting agent, (iii) infeasible for an entrusting agent to compute to make computations of whether the resources saved by having other agents (other than the entrusted agent) perform certain actions outweighs the resources saves by having the entrusting agent perform them, and (iv) infeasible to compute the world contingencies that might happen that need to be addressed by those agents (other than the entrusted and entrusting agents) who perform those actions that are necessary for the performing of the action that the entrusted agent is normatively expected to perform. That all of these computations are infeasible means that there is an iterative structure to the trust relation, which we will describe, and which we will argue is the justification for why resource conservation is a necessary condition of a trust relation.

The Buechner-Tavani model of moral trust

Buechner and Tavani, combining the views of Strawson, Holton, and Walker, propose a relationship of trust between *A* and *B* that is one in which the following five conditions obtain:

A has a normative expectation (which may be based on a reason or motive) that *B* will do such-and-such;

B is responsible for what it is that *A* normatively expects her to do;

A has the disposition to normatively expect that *B* will do such and such responsibly;
A's normative expectation that *B* will do such-and-such can be mistaken;
[Subsequent to the satisfaction of Conditions (i) – (iv)] *A* develops a disposition to trust *B*.[10]

That *A* has the disposition to trust *B* to do such-and-such allows room for cases in which *B* does not do such-and-such, even though *A* trusts that *B* will do it. Condition (iv) allows for cases in which *A* mistakenly trusts *B*. There may be over-determined cases in which, for instance, *A* expects *B* to do such-and-such both because *A* trusts *B* and because *A* has been coerced into expecting *B* to do such-and-such. As long as *A* normatively expects *B* to do such-and-such, and it is not the case that, for instance, the normative expectation is itself the result of being coerced, there is a genuine trust relationship in place (provided the other conditions are satisfied as well). The definition rules out as instances of trust those cases in which *B* has a responsibility to do such-and-such, even though *A* does not attribute responsibility to do such-and-such to *B*.

A normative expectation is in place when "I rely on you to do what you should. I do not only expect *that* you will do it, I expect it *of* you."[11] The connection between an expectation and a normative expectation connection is not a logical connection—in either direction it can come apart. Here is how it can come apart. (i) Expectation does not imply normative expectation. A can predict that B will perform some action, but not normatively expect that B will do it. (ii) Normative expectation does not imply expectation. *A* can normatively expect that *B* will do such-and-such, even when *A* does not expect (predict) that *B* will do it. That is, either *A* does not predict that *B* will do it (the vacuous case) or A predicts that B will not do it, but still normatively expects that B will do it. This shows that a normative expectation and an expectation are really two different concepts. It might be thought that where A normatively expects that B will perform an action, that implies that A expects that B will perform that action. But that is so only if a normative expectation is a sub-concept of the concept of expectation. It would be better to distinguish a normative expectation from a predictive expectation. Each is a kind of expectation.

Under what conditions does a normative expectation arise? To say that I expect something—say, doing X—of you is to say that you will act as you should. But that means that I "embody a strong anticipation of due care and compliance with [normative] standards—in other words, of responsibility—from others and ourselves."[12] Given that you will act as you should, in normatively expecting you to do so, one is entitled to your acting in that way. If you do not act in

that way, and it is not an excusable omission, I can rightfully express a reactive emotion—such as resentment—towards you. Thus, much is morally at stake in a relation of trust. But there is more—for the world is a complex place in which anticipated actions can go wrong in unanticipated ways. Surely, a responsible trustee should have some awareness of such world contingencies that might thwart her actions that the trustor entrusts her to perform.

Given (i) the contingencies of the world that might conspire to thwart you doing what you should do, and that (ii) both of us know that these contingencies exist, and that (iii) it is common knowledge that they exist, we have good reason to act in ways that either anticipate the contingencies or else recognize them when they arise. In trusting you to do A, I also entrust you to parallel actions of anticipation and recognition with respect to the world contingencies. Moreover, in normatively expecting you to do that, I am placed under a similar entrustment on your part—to be responsible for parallel actions of anticipation and recognition with respect to world contingencies. Thus, the additional condition on trust is a reciprocal relation of trust with respect to the kind of contingencies that might defeat any trust relation. There are options—we might take the reciprocal relation to be a proto-trust relation, rather than an additional condition on the trust relation. Or we might take the reciprocal relation to be a kind of insurance policy, which normatively binds those who agree to it. However, which option we choose depends upon the nature of the justification provided for the reciprocal relation.

But, it might be objected, how is an additional condition that is levied on the conception of trust justified? It does no good to simply stipulate that the conception of trust must have this additional condition. There must be good reasons—and reasons internal to the conception of trust, that justify the addition. We will argue that the justification of the condition that both trustor and trustee are aware of world contingencies and that the trustee addresses, as part of her responsibility toward the trustor, those world contingencies is in terms of the features of the trust relation we have already delineated—namely, that it conserves resources. Indeed—that it conserves the resources of the trustee and the resources of the trustor.

Iteration, reinforcement, and bootstrapping of the moral trust relation

Where a normative expectation has the form "I rely on you to do what you should. I do not only expect *that* you will do it, I expect it *of* you" in the context of addressing world contingencies, it gets iterated in the sense that it applies to any contingency that might arise that is relevant to maintaining the moral trust relation. The iteration occurs in two distinct ways. The

first way the iteration occurs is across a temporal sequence of world contingencies (which is either anticipated or not anticipated by a trustor). The trustor might predict a world contingency at t_1, a world contingency at t_2, and a world contingency at t_3. She would then hold B responsible for addressing that temporal sequence of world contingencies. The second way in which the iteration occurs is where one world contingency causally brings about another world contingency. Suppose that in addressing world contingency 1, world contingency 2 is causally brought about. In that case, A normatively expects that B will take care of world contingency 2. Of course, both ways are not mutually exclusive. Suppose there are two world contingencies that arise in the context of a trust relation between A and B, for B to do X. Then the iteration of normative expectations means that A normatively expects B to take care of world contingency 1 and to take care of world contingency 2. But there might be another level to the iteration, where attending to world contingency 2 causally brings about world contingency 3. Then B will attend to world contingency 3 as well.

But notice that there is a problem here. Clearly, it *might* happen that the chain of independent world contingencies, or the chain of dependent world contingencies, reaches some length that is beyond the capabilities of any human actor to address. Indeed, no human being could know, ahead of time, that this is not the case—no human being could know ahead of time the exact length of the chains and that the chain lengths do not exceed a length that can be humanly addressed. However, given that no human being could have this knowledge, it would be rational for any human being to temper, in some way, the expectation that the trustee will, in fact, address the world contingencies as they arise. One way to do this would be to add a proviso to the expectation, perhaps of the form: "I do not expect that the trustee will have the capacity to address any arbitrary chain of world contingencies, and so will not expect that he is able to address any arbitrary chain of world contingencies." This formulation is certainly how a rational agent would reformulate their expectations concerning how a trustee addresses world contingencies. But it is important to remember that in the context of a trust relation we have more than a mere extensional expectation, We have the resource of a normative expectation, which provides an expectation with a normative component, in such a way, that there is no longer a necessary or sufficient condition connection (in either direction) between expectations and normative expectations. This resource allows us to formulate a normative expectation about world contingencies in terms of how a trustee will be responsible for addressing them.

Certainly, it would be absurd to formulate such a normative expectation too stringently. We could not hold an agent who is a trustee to be responsible for doing something which she does not have the capacity to do. Such absurdities surface in various systems of modal deontic logic, and in those cases it is essential that they be eliminated. Similarly, a morality that is not livable—whose tenets could not be satisfied by human agents—would be an unworkable moral system. The question, then, is how to make a normative expectation about world contingencies that need to be addressed by a trustee one that is humanly livable. Since it is impossible for any human actor to know ahead of time when an arbitrary chain of world contingencies is humanly addressible and when it is not, the obvious way out of the impasse is the following: the trustor should *trust* that the trustee address world contingencies as they arise. But where the chain of world contingencies exceeds the capacities of a human agent to address them, the trustor knows that the trustee will not be able to address them. The trustor trusts that the trustee will address those world contingencies that fall within her capacity to address, and where they do not, the trustee will not be able to address them. The trustor does not expect, *of* the trustee, that she be able to address world contingencies that exceed her capacities. That is, the trustor does not expect, *of* the trustee, that she do something which she (or any other human agent) cannot do.

But this solution raises a troubling question: Isn't it the case that where a trusting agent trusts that the trustee will address world contingencies, the definition of a trust relation cannot accommodate this condition without becoming either circular or being committed to an infinite descending chain of trust conditions? That is, introducing into the definition of a trust relation that a trusting agent trusts a trustee to do such-and-such presupposes that there is an antecedent definition of trust. If so, then there would have to be an antecedent definition of trust before one could define a trust relation! Surely, one should avoid such logically suspect constructions in defining a trust relation!

The key to solving this logical priority problem is to see that it is a necessary feature of a trust relation that it conserve resources, and that where a trust relation is invoked in the definition of a trust relation in order to function as a means of conserving resources, it is not circular to introduce it into the definition. But if conservation of resources is *not* a necessary condition on a trust relation, then the logical priority problem remains unsolved. One can take this to be a justification of the claim that it is a necessary condition in the definition of a trust relation that moral trust relations conserve resources.

Thus, the iteration of the normative expectations reinforces the moral trust relation. But there is also a bootstrapping effect as well—the very things that need to be done to address world contingencies might become actions which one agent entrusts to another agent. This need not happen—and it is not a necessary part of the trust relation that it does happen. So bootstrapping is a contingent feature of the moral trust relation. Describing the formal structure of iteration, reinforcement, and bootstrapping is a relatively simple matter, but for reasons of space it will not be done here.

Some misconceptions about trust and resource conservation

One objection to the claim that resource conservation is a necessary part of the moral trust relation is the following: "If I am walking down the street, and I am not actively thinking about all the bad things people might do to me, you seem to suggest that I am trusting people, in part to conserve the resources necessary to evaluate each potential threat."[13]

First, we need to distinguish (i) walking down the street, not thinking about the bad things which other people might do to me from (ii) walking down the street not thinking about the bad things which other people might do to me *because* I trust that other people will not do those bad things to me. Clearly, (i) might be true even though (ii) is false. That is, (i) can be true even though there is no trust relation in place between the subject doing the walking and other people who might harm him in one way or another.

Let's suppose that the content of the trust relation is that I trust other people—call a generic representative of them OP—not to harm me when we are in a social communal setting. Then "I am not actively thinking about the bad things people will do to me" is not sufficient for that trust relation to be in place. Rather, in trusting people not to harm me, I will not actively think about the bad things they might do to me. That is, each potential threat to my safety is not actively evaluated unless something is out of the ordinary.

This raises two important questions. The first question is: what are the normal conditions for a merely potential threat to me by OP—the conditions under which there is no need to actively evaluate that threat. We first have to define what we mean by a "merely potential threat," and distinguish it from active threats. We also have to know the conditions under which some action constitutes a threat, under which there are the possibility of threats, but no actions that are threatening, the difference between the normal conditions under which an action is a threat, and the abnormal conditions under which a non-threatening action is taken to be a threat.

Worse yet, there are second-order abnormal conditions concerning abnormal conditions under which a non-threatening action is taken to be a threat, under which it is not a threat. Notice that it would be infeasible for an entrusting agent to make these computations and distinctions. Indeed, even for an agent not engaged in a trust relation, it would be infeasible to make them. This shows that it is part of the trust relation to dispense with having to make the computations and distinctions.

The second question is: How does it happen that "I will not actively think about the bad things they might do to me?" Does this condition necessarily follow establishing a trust relation, precede establishing a trust relation, or is it part of the establishment of a trust relation—in the sense that there is a formal, interactive structure between agents? By itself, it is underdetermined. However, once we know that an agent is engaged in a trust relation with other OP, it follows that it is motivated by that trust relation, and that it follows from being in that trust relation.

A moral trust relation is not wholly a matter of not actively thinking about how other people might harm one (where one trusts that OP will not harm one). Not actively thinking about what harms might come from other people is neither a necessary nor a sufficient condition for a moral trust relation. There will be conditions under which an agent must actively think about harms others might perpetrate upon him. This shows that it is not a necessary condition. There are also normative necessary conditions which must be in place for a moral trust relation. Recall the normative features of trust in the Buechner-Tavani model of moral trust. All of these normative features must be in place for a relation to be an instance of a moral trust relation. This shows that it is not a sufficient condition. However, different situations in which a moral trust relation arises will present complex contingencies that it is part of that relation to relegate to the background, in the sense that there is no need to explicitly consider and act upon the various problems that the situation and the trust relation generate. Indeed, it is part of the trust relation that these problems will be taken care of as they arise—either explicitly or by just letting things alone.

Of course, it could be claimed that there is an independent mechanism at work which takes care of these frame and frame-like problems. One suggestion is that there is an independent epistemic mechanism whose basic structure takes the form: "I simply chose to ignore any potential threats." The problem for this prescription is that it is all too easy to simply say that I will ignore any potential threats. But how do you choose to ignore them—by enumerating a list

of potential threats and then attending to each item on the list, perhaps marking it with the words "I will not attend to this item!"? Notice that this independent mechanism relies on a distinction between accidental and intentional ignorance. However, my point is that intentional ignorance is incoherent in the absence of some means of achieving reliability—of determining when one must stop being intentionally ignorant and attend to something-at-hand. Intentional ignorance in n-person game situations—in which each player recognizes that all of the other players might be able to determine when to attend to something-at-hand and that this is common knowledge among all of the players, is one way of describing the way in which a trust relation conserves resources.

One way to see how the idea of intentional ignorance requires an independent mechanism for discerning abnormalities is apparent in the following (*modus tollens*) inference:

> If things were out of the ordinary, then I would know it.
>
> I do not know thing are out of the ordinary.
>
> Therefore, things are not out of the ordinary.

Unless there is a detection mechanism for what is out of the ordinary—which would provide a good reason to think that the first premise is true—there is no reason to conclude that things are not out of the ordinary.[14] However, if a trust relation is in place, then an agent might not have to exhibit a form of intentional ignorance. Rather, she can exhibit a form of accidental ignorance, since it is part of the trust relation that one trusts that world contingencies are addressed as they occur by the entrusted agent. An entrusting agent can be accidentally ignorant of the ways in which world contingencies can disrupt the actions the entrusted agent is to perform, since she trusts that they will be addressed when they arise. There is no need to be in a state of intentional ignorance, in which one makes inferences of the form "If things were out of the ordinary, then I would know about it." This is yet another way in which the trust relation conserves resources. It raises an important question, however. Namely, in determining whether an agent is trustworthy, will an agent have to make all of the computations which it is the point of a trust relation to avoid making? In particular, in determining that an agent is trustworthy, does one have to first determine the conditions under which things might be out of the ordinary in that agents detection of world contingencies? If an entrusting agent does not establish the trustworthiness of a agent to whom she intends to entrust the performance of certain actions, then she cannot be justified in thinking that agent will, if entrusted, perform those actions. In which case, if justification of an entrusted agent's trustworthiness is a necessary feature of entrusting an agent with performing

some action, then these inferences and computations will have to be made before there is adequate justification.[15]

Knowing when things go wrong and (some implications for) multiple agents

When it comes to determining when world contingencies need to be addressed, such as that something is badly out of the ordinary, two heads are typically better than one. In general, *n* heads are typically better than *n-1* heads. A reason for agent *A* to trust agent *B* is that not only will *B* assume responsibility for doing something of salience to *A*, but will also assume responsibility for addressing world contingencies which might thwart doing that. Moreover, *B* knows that *A* will also assume responsibility for addressing world contingencies. That two heads are better than one is a reason to engage in a trust relation where agents know that world contingencies could defeat plans of any kind to engage in various kinds of actions. For a specific example, recall Baier's claim that in order for *B* to care about something that *A* cares about, both A and B must exhibit a certain amount of discretion. It is not feasible to enumerate the kinds of situations relevant to caring about something and the various ways in which discretion manifests itself in those situations. Rather, *A* has a good reason to trust *B*, and *B* has a good reason to trust *A*, to see to it that such contingencies are addressed as they arise, since both *A* and *B* know that two heads are better than one when it comes to addressing those contingencies. Obviously, *A* can help *B* by telling *B* about what *A* cares about, and how that caring plays itself out in various contexts. *B* can help *A* by pointing out how someone other than *A* can interpret those contexts, and the caring that is involved, in ways not anticipated by *A*. Knowing that two heads are better than one provides a reason, within the concept of moral trust, for *A* and *B* to engage in the moral trust relation.

But it would be a world in which we simply do not live if it were the case that both *A* and *B* act independently of everyone else in that world. It might be called a Robinson Crusoe world, since in order to act independently of everyone else, there must not be anyone else in that world. (Strictly speaking, even Robinson Crusoe trusts others—that, for instance, he was properly taught arithmetic in grade school so that when he surveys his land on the desert island on which he is stranded, he gets the right measurements.) In any trust relation between two principals, *A* and *B*, other trust relations obtain between *A* and others (one of which might be *B*), and between *B* and others (one of which might be *A*). It is the social setting in which individuals are situated which provides means and opportunities for activities. A human action is not a solitary event, but

depends on the social setting in various ways, some of which are obvious and known to the actors, and some of which are not known to the actors. In either case, there is a trust relation between the various actors in that social setting.

Buechner and Tavani have explored the trust relation in such social settings, which they have called the diffuse, default model of trust.[16] Given the social setting in which all actions occur, it is the case that when *A* trusts *B* to perform some action, *A* must also trust countless others that the environment remains stable enough for that action to be performed. Similarly, *B* must also trust countless others that the environment remains stable so that she can perform that action. There is, however, an additional point. It is that A knows and that B knows that that action can be performed, simply because each of the countless others upon whom they depend know that that the action can be performed. Or—to be precise, each of the countless others is justified in believing that some piece of the social scenario in which the action can be successfully performed is in place. In this way, A and B have social knowledge. Their knowledge that the environment is stable, and that the action can be successfully performed (given that b does successfully perform it) is social knowledge, since it depends upon the fact that countless others each have justified beliefs that the various components of the social setting are in place (i.e., the components which are necessary for the action to be performed).

That there is such a social setting for all human actions reveals an interesting structural relation in the concept of moral trust. I have already shown that addressing world contingencies is part of the conceptual structure of a moral trust relation. That I know the world contingencies will be addressed follows from immersion in a moral trust relation. But given the social setting in which the actions performed and world contingencies occur, I must trust that those others who will address those aspects of the action and world contingencies that are necessary for establishing a stable social environment in which the action can successfully be performed, Indeed, I know that those aspects will be addressed, for in trusting that others get it right, I rely on an epistemic trust. That is, I know they will get it right because I trust that they will get it right. But this social knowledge must be in place before a moral trust relation between two individual actors can be established. Moral trust depends on epistemic trust, because a moral trust relation is undertaken in a social setting, where countless social actors are necessary for actions to take place because the actions depend on a stable social environment and it is the

individual functional roles of the countless social actors who ensure that the social environment is stable.

Frame problems, moral trust, and epistemic trust

There are other normative areas in which frame problems arise—in particular, frame problems arise for any theory of epistemic justification. It is an ingenious insight of Kent Bach to see that a doxastic habit that would epistemically justify both basic and conditional everyday beliefs consists in the application of default reasoning in inferring those beliefs, where explicitly reasoning to acquire those beliefs would be computationally infeasible.[17] How does default reasoning solve the resource limitation problem that explicit reasoning encounters? In default reasoning, steps are taken by default. That is, one jumps to a conclusion unless there is a reason not to jump to that conclusion. The important feature of default reasoning, as Bach conceives of it, is that it is not merely jumping to a conclusion, period. If default reasoning were the kind of reasoning, then it would hardly qualify as a means of conferring epistemic justification upon the beliefs at which one arrives using it. For, jumping to a conclusion without considering any kind of evidence for that conclusion is no better than arriving at a belief on the basis of flipping a coin. Without considering any evidence for the belief, we have no more reason to believe that it is true than we have reason to believe that it is false. Rather, default reasoning, in order to confer epistemic justification upon the beliefs that are inferred using it, must contain provisos about what happens when conditions are not ordinary.

Bach formulates what he dubs the take-for-granted principle (TFGP) in the following way: "(TFGP) Its appearing to one that p justifies directly inferring that p provided that (a) it does not occur to one that the situation might be out of the ordinary, and (b) if the situation were out of the ordinary, it probably would occur to one that the situation might be out of the ordinary."[18] He claims—though without any proof—that beliefs that are arrived at on the basis of TFGP are reliable; that is, such beliefs are epistemically justified.

The resource conservation idea is that there are inferential steps in using TFGP that are not explicitly taken, viz., the steps that are described by conditions (a) and (b) in the formulation of TFGP. It is in not having a thought that things might be out of the ordinary and in the truth of the counterfactual, that if things were out of the ordinary, then it would occur to one that they are out of the ordinary, that one avoids having to make explicit inferences. Rather, one simply infers p, period. In inferring that p is the case, the conditions (a) and (b) are not explicitly followed.

Conditions (a) and (b do not come explicitly to mind in making the inference that p is the case. These conditions are not explicitly expressed or represented in one's reasoning; rather they are implicit in the reasoning. But what does that mean?

Detection mechanisms and when they should be activated

Conceding that there is an adequate notion of what it is for a piece of reasoning to implicitly assume something, there are additional problems to which Bach's definition of TFGP succumbs. He says that "a belief resulting from such a process [TFGP] is justified to the extent that the process not only leads to true beliefs, but also guards against forming false beliefs, by means of precautionary subroutines that are generally activated when and only when they need to be. For it only to that extent that following TFG[P] can lead to justified beliefs."[19] The problem is that it is not clear what he means by the phrase "when and only when they need to be." First, is he taking 'when and only when' to mean logical equivalence? If so, and if it means that the precautionary subroutine is activated when and only when one would, if it were not activated, infer a false belief, then it is much too strong. It would make the use of default reasoning infallible. That might be quite nice, but it is hardly realistic. What sort of detection mechanism must one have in place in order to be infallible in inferring one's perceptual beliefs by means of default reasoning? It would have to be a detection mechanism that infallibly detected each and every abnormal situation in which one would, if it were not activated, infer a false perceptual belief. On the other, hand, if the notion is loosened up a bit, how much loosening is permitted? That is, how often can one fail to engage a precautionary subroutine (and thus infer a false belief) and still be reliable in inferring true perceptual beliefs? This is a difficult problem that will not go away.

What Bach did not notice is that the problem of providing a detection mechanism will be easier when there are two or more agents than when there is a single agent. Epistemic trust—coupled with moral trust—will provide a detection mechanism which depends upon (i) the innate detection mechanisms of two (or more) agents and (ii) the iterative, bootstrapping moral trust relation. Thus, what is added to the innate detection mechanisms is a responsibility for detecting abnormalities if and when they occur. That is, it may well be the case that social knowledge is a prerequisite for individual knowledge, in the following way. Detection mechanisms may be easier to come by when two or more agents are involved than where there is merely a single agent involved. Hence, epistemic reliability will be easier to come by when there are two or more

agents. Obviously, this could not be credible in the case of perceptual knowledge, since that is knowledge with respect to a solitary individual. Unless one is an extreme behaviorist, perceptual knowledge is individualistic. It would be too far afield to go into at this point, but it is clear that there are connections between epistemic trust, moral trust, and the epistemic justification of individual knowledge.[20]

Trust and *Ecological Rationality*

'Ecological rationality' names a fairly broad project which Gerd Gigerenzer and his colleagues have been working in since the early 1980's: it is basically the view that "we are able to achieve intelligence in the world by using simple heuristics in appropriate contexts."[21] Behavior that adapts to a particular kind of environment is intelligent when the appropriate kind of interaction between mind and that environment occurs. This interaction, to be successful, requires that there is a match between the informational structure of that environment and the informational structure—or the information processing mechanisms—of the heuristics a human being employs in navigating that environment. The utility of heuristics is that they conserve agent resources. In uncertain environments, the amount of data and computation that would be needed to ensure that a given action succeeds with respect to normative standards will be infeasible in many cases.

Ecological rationality provides analyses of both the informational mechanisms of the human mind and the informational structure of the environment, with an eye toward revealing the consonance of the two. Prior to this work, rationality was taken to be a normative notion informed by logic and by decision-theory. Behavioral departures from the constitutive norms of rationality were taken to be irrational behaviors. But Gigerenzer and his colleagues have argued that such behaviors need not be irrational. If the behaviors are guided by heuristics that are consonant with the informational structure of the environments in which they are applied, then the behavior is rational. Under ecological rationality, heuristics can be normative, and not merely descriptive of behaviors that are informed by them.

Given that a necessary feature of a moral trust relation is to conserve agent resources, there is a natural connection between moral trust and ecological rationality. One might consider a analogy between trusting that the environment will remain stable so that one's actions will succeed and heuristics that are employed in acting in uncertain environments where one does not know if it will remain stable over some temporal interval. If so, then there are two claims which

can be defended. The first is that trust is a mechanism by which we can achieve ecological rationality. The second is that conservation of resources is a necessary condition on a trust relation. It follows from these two claims that trust is necessarily constitutive of ecological rationality. Notice that trust would adapt to almost any environment, and that there would be no need for evolutionary trials of specific kinds of heuristics to winnow out those heuristics that fail in environments of a given type and to select those heuristics that succeed in environments of a given type. The point, though, is that taking conservation of resources to be a necessary condition on a moral trust relation—that conservation of resources is constitutive of the concept of moral trust—allows the concept of moral trust to be applied to new areas which, *prima facie*, appear quite alien. The apparent incongruence dissolves and new work for the concept of moral trust opens.

References

1. McCarthy, J, Hayes, P. Some Philosophical Problems from the Standpoint of Artificial Intelligence in Meltzer, B, Mitchie, D. (eds.) *Machine Intelligence 4*; Edinburgh University Press: Edinburgh, UK, 1969.

2. Baier, A. Trust and Anti-Trust. *Ethics* **1986,** 96(2), 231-260.

3. Walker, M. *Moral Repair*, Cambridge University Press: New York, USA, 2006.

4. Walker, M. op. cit., p. 78.

5. Govier, T. *Trust and Human Communities*; McGill-Queen's University Press: Montreal, CAN, 1997.

6. Jones, K. Trust as an Affective Attitude. *Ethics* **1996,** 107(1), 4-25.

7. Baker, J. Trust and Rationality. *Pacific Philosophical Quarterly* **1987,** 68(1), 1-13.

8. Shoham, Y. *Reasoning About Change: Time and Causation from the Standpoint of Artificial Intelligence*, MIT Press: Cambridge, USA, 1988.

9. Shoham, Y. op. cit.

10. Buechner, J, Tavani, H. Trust and Multi-Agent Systems: Applying the "Diffuse, Default Model" of Trust to Experiments Involving Artificial Agents. *Ethics and information Technology* **2011,** 13(1), 39-51.

11. Walker, M. op. cit. p. 67.

12. Walker, M. op. cit. p.68.

13. Remarks made by a reviewer of an earlier version of this manuscript for the CEPE 2011 Conference.

14. Moore, R. Semantical Considerations on Non-monotonic Logic. *Artificial Intelligence* **1985,** 25(1), 75-94.

15. I leave this important matter for another paper, since it would require too much space to argue the position here. See Buechner, J, Simon, J, Tavani, H. Re-Thinking Trust and Trustworthiness in Digital Environments, forthcoming.

16. Buechner, J, Tavani, H. op. cit.

17. Bach, K. A Rationale for Reliabilism. *Monist* **1985, 68(2),** 246-263.

18. Bach, K. op. cit.

19. Bach, K. op. cit.

20. I explore these connections in detail in a forthcoming paper. See Buechner, J. Moral Trust and Epistemic Trust as Epistemological Concepts, forthcoming.

21. Gigerenzer, G, Todd, P. *Ecological Rationality*, Oxford University Press: New York, USA, 2012.